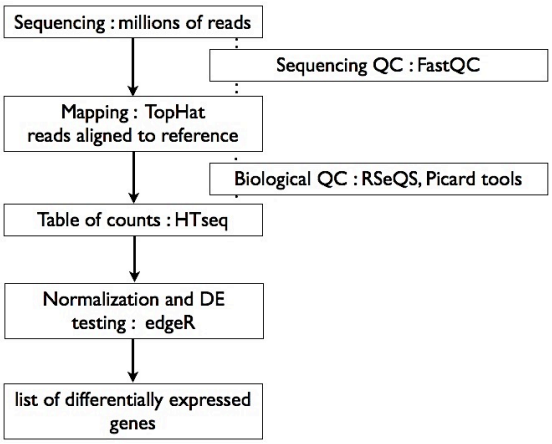
University of Geneva – iGE3 Genomics Platform – CMU
O. Stojanovic, M.Trajkovski, Project : Ob-jej
July 2016 - RNASeq analysis

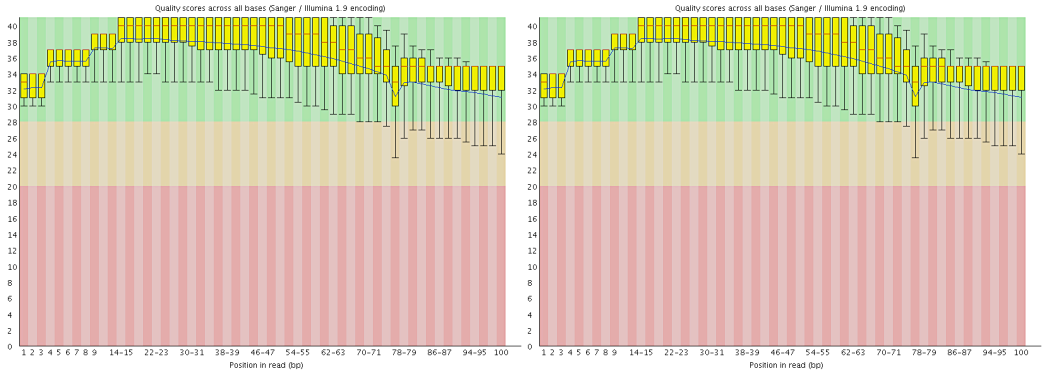
RNAseq sequencing and analysis pipeline used:



Sequencing and quality control

Sequencing: single read 100, TruSeq stranded, ribodepletion

The sequencing quality control was done with FastQC and is good for all samples. The quality distribution along the reads plot is shown for ob1 (left) and wt2 (right) samples.



Note: quality scores of around 30 and above correspond to 1/1000 chance of errors and a quality score of 40 to 1/10'000 chance of errors.

The reads raw fastq files are available trough the LIMS (http://uhts-gva.vital-it.ch/lims).

The following table gives the number of raw reads sequenced.

Library	Run	GAP	Lane Yield	Raw Clusters#	% PF Clusters	% Q30 PF	Mean Score PF
ob1	SN865 #562 L4	1.82	3,362	33,615,525	100	90.95	35.63
ob2	SN865 #562 L4	1.82	3,584	35,844,916	100	90.82	35.58
ob3	SN865 #562 L4	1.82	3,529	35,287,952	100	90.87	35.6
wt1	SN865 #562 L4	1.82	3,612	36,122,290	100	90.75	35.57
wt2	SN865 #562 L4	1.82	3,219	32,186,153	100	90.81	35.59
wt3	SN865 #562 L4	1.82	3,815	38,150,655	100	90.96	35.63

Mapping and biological quality control

The reads were mapped with the **TopHat v.2** software to the UCSC mm10 reference; on new junctions and known junctions annotations. Biological quality control and summarization were done with **PicardTools1.92**.

The alignment bam files are provided upon request. The average mapping rate is 93.27%.

Sample	#input reads	#mapped reads	pc multiple matching reads	pc overall mapping
ob1	33'615'525	31'425'622	9.40%	93.50%
ob2	35'844'916	33'523'118	9.00%	93.50%
ob3	35'287'952	32'868'331	8.80%	93.10%
wt1	36'122'290	33'633'284	8.60%	93.10%
wt2	32'186'153	29'833'401	8.70%	92.70%
wt3	38'150'655	35'743'661	9.00%	93.70%

The picard tools RNAseq quality metrics are given in the following tables. The percentage of mRNA bases is on average 52.2%. The other half of the reads fall into intergenic or intronic portions 47.52%.

Sample	wt1	wt2	wt3	ob1	ob2	ob3
PF_BASES	3'363'328'400	2'983'340'100	3'574'366'100	3'142'562'200	3'352'311'800	3'286'833'100
PF_ALIGNED_BASES	3'363'171'731	2'983'204'156	3'574'213'856	3'142'427'991	3'352'170'635	3'286'690'896
RIBOSOMAL_BASES	16'302'500	7'506'800	7'275'500	5'883'200	7'929'000	7'618'800
CODING_BASES	1'134'699'961	990'175'764	1'245'236'060	1'126'770'840	1'210'700'716	1'178'976'919
UTR_BASES	566'884'576	496'443'304	614'622'642	560'111'723	592'067'684	575'115'372
INTRONIC_BASES	1'149'494'233	1'030'171'142	1'173'223'129	984'059'625	1'052'926'958	1'052'011'862
INTERGENIC_BASES	495'808'401	458'919'893	533'867'181	465'607'624	488'554'184	472'976'318
IGNORED_READS	0	0	0	0	0	0
CORRECT_STRAND_READS	16'324'357	14'270'159	17'797'467	16'095'597	17'196'552	16'763'407
INCORRECT_STRAND_READS	127'096	117'684	132'167	114'005	117'949	119'766
PCT_RIBOSOMAL_BASES	0.48%	0.25%	0.20%	0.19%	0.24%	0.23%
PCT_CODING_BASES	33.74%	33.19%	34.84%	35.86%	36.12%	35.87%
PCT_UTR_BASES	16.86%	16.64%	17.20%	17.82%	17.66%	17.50%
PCT_INTRONIC_BASES	34.18%	34.53%	32.82%	31.32%	31.41%	32.01%
PCT_INTERGENIC_BASES	14.74%	15.38%	14.94%	14.82%	14.57%	14.39%
PCT_MRNA_BASES	50.59%	49.83%	52.04%	53.68%	53.78%	53.37%
PCT_USABLE_BASES	50.59%	49.83%	52.03%	53.68%	53.78%	53.37%
PCT_CORRECT_STRAND_READS	99.23%	99.18%	99.26%	99.30%	99.32%	99.29%
MEDIAN_CV_COVERAGE	0.465164	0.470846	0.468314	0.471942	0.472025	0.468992
MEDIAN_5PRIME_BIAS	0.085525	0.085219	0.086785	0.085747	0.083692	0.090013
MEDIAN_3PRIME_BIAS	0.736095	0.728826	0.754494	0.726981	0.722121	0.722481
MEDIAN_5PRIME_TO_3PRIME_BIAS	0.161216	0.154448	0.152995	0.162104	0.166528	0.176976

Table of counts:

The table of counts with the number of reads mapping to each gene feature of UCSC mm10 reference was prepared with **HTSeq v0.6p1** (htseq-count).

The table of counts before normalization is provided in excel format.

File : raw_counts.xlsx

Differential expression analysis and results:

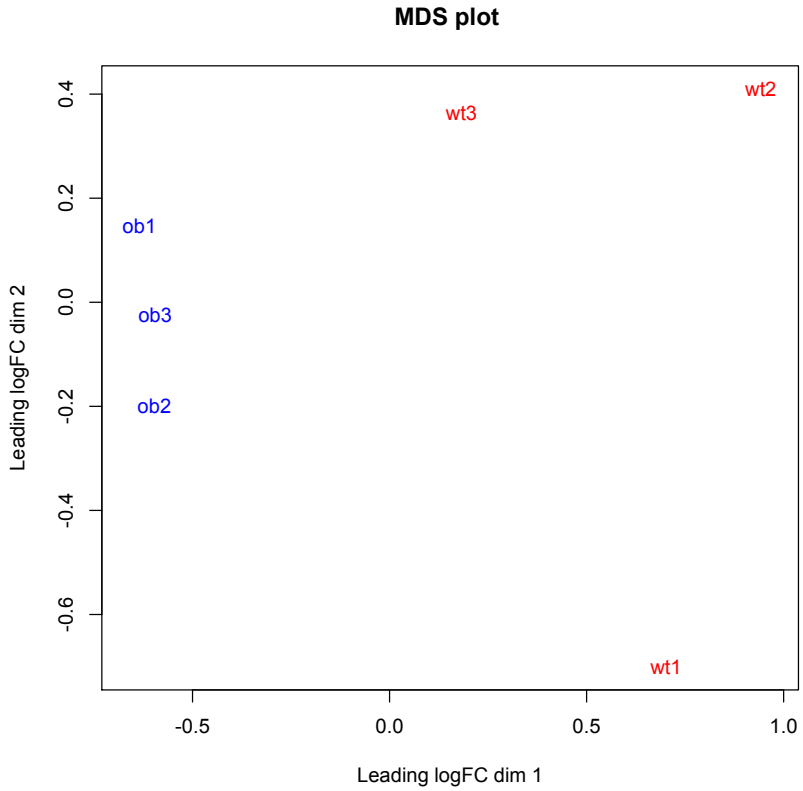
The differential expression analysis was performed with the statistical analysis R/Bioconductor package **EdgeR v. 3.4.2,** for the genes annotated in mm10. Briefly, the counts were normalized according to the library size and filtered. The genes having a count above 0.6 count per million reads (cpm), i.e. corresponding to 10 counts, in at least 3

samples were kept for the analysis. The differentially expressed genes tests were done with t-test, negative binomial distribution.

The normalization factor and library size are shown in the following table.

	group	lib.size	norm.factors
ob1	**ob**	16'599'355	0.98
ob2	**ob**	17'761'864	0.93
ob3	**ob**	17'250'503	0.97
wt1	**wt**	16'805'216	1.06
wt2	**wt**	14'661'981	1.04
wt3	**wt**	18'286'080	1.02

Multidimensional scaling MDS plots of the samples is shown below. The MDS plot gives an indication of the similarity, based on the fold changes between all the pairs of samples.



The samples separate well according to the treatment.

The BCV biological coefficient of variation is a value computed in the edgeR analysis that indicates how much the samples counts are different. In this data set, the BCV is 0.095 (10% of variation in gene expression between samples can be expected).

Differential expression analysis:

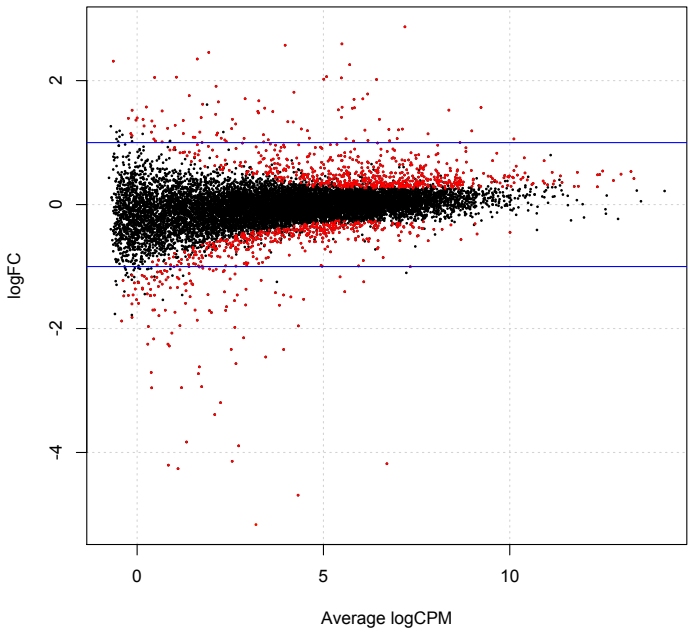
The raw gene number of the set is 23’420. The poorly or not expressed genes were filtered out. The genes expressed with least at 0.6 cpm (count per million reads) in 3 samples are kept. The filtered data set consists of 13’969 genes.

The differentially expressed genes p-values are corrected for multiple testing error with a 5% FDR (false discovery rate). The correction used is Benjamini-Hochberg (BH).

The following table gives the differentially expressed genes statistics (FDR 5%) and the number of which have a fold change 2 threshold.

	Significantly differentially expressed genes FDR 5%					
Test	**# down regulated**	**# no change**	**#up regulated**	**#FC 2**	**< 2**	**> 2**
Ob versus WT	523	12'730	716	212	121	91

MA plots of the differentially expressed genes. Significantly differentially expressed genes are shown in red. The blue line shows the FC 2 and -2 (log2 FC 1).



Note :
The logCPM (count per million) value shown in the differentially expressed genes output files, is the average log2 count per million for a gene taken over all the libraries in the dataset. It is a measure of the overall expression level of a transcript and is displayed in the dispersion and biological variation plot above. It enables to weight the fold change according to the expression of the gene in the data set.

The output files provided are:

1). Fold changes for all the genes in the set.
3). Differentially expressed genes with a fold change 2 threshold and a FDR 5% correction.
4). Raw counts.
5). Normalized counts.

References:

FastQC : http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

TopHat : Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* (2009) 25 (9) : 11005-1111

Picard Tools : http://picard.sourceforge.net/

HTseq : http://www-huber.embl.de/users/anders/HTSeq/

edgeR: Robinson MD, McCarthy DJ and Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139-140